nature portfolio

Peer Review File

Aberrant Cellular Communities Define Disease Heterogeneity in Chronic Obstructive Pulmonary Disease

Corresponding Author: Dr Maor Sauler

Version 0:

Decision Letter:

16th Dec 2024

Dear Dr Sauler,

Thank you for submitting your manuscript entitled "Aberrant Cellular Communities Define Disease Heterogeneity in Chronic Obstructive Pulmonary Disease". We have given the paper our careful consideration and find it of potential interest. However, due to certain shortcomings we are concerned that sending the current manuscript out to review could lead to unnecessary delays and quite possibly an undesirable outcome of the review process.

Briefly: your work does not contain sufficient information on the data QC, for all the omics data but especially for the snRNA.

We know from extensive experience that reviewers request substantial detail on the QC of such single-nucleus/-cell data, to ensure that your results are not confounded by technical differences associated with e.g. the disease state of the sample, etc.

The lack of such important information would likely lead to an unconstructive first round of review, with referees saying it's impossible to judge the veracity of results.

We would therefore ask that you supply full QC on your data. As a non-authoritative example, we think https://github.com/hbctraining/scRNA-seq_online/blob/master/lessons/04_SC_quality_control.md is a good start, but we would highly recommend you and your team also look further around to ensure you are following field standards.

We would also ask that you consider the same question for the other (proteomics, Olink, GeoMX) data you present with the snRNA.

We would therefore like to invite you to revise your manuscript to address these concerns before we make a final determination on whether to send your manuscript for external review.

We shall hope to receive your revised version as soon as you are able to complete the suggested revisions. If something similar is published in the interim we will have to consider the impact it has on the novelty of a revised manuscript.

If you anticipate a delay of more than four weeks, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Genetics or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date may be that of the revised, not the original, version.

If you are not interested in submitting a suitably revised manuscript in the future please let me know immediately so we can close your file. If you have any questions, please contact me.

1) Please ensure that you have completed the Reporting Summary required for review:
https://www.nature.com/documents/nr-reporting-summary.pdf

2) Please also complete the Editorial Policy Checklist that would be a requirement for eventual publication in a Nature journal:
https://www.nature.com/documents/nr-editorial-policy-checklist.pdf

Please be aware of our guidelines on digital image standards.

Please use the link below when you are prepared to resubmit.
Link Redacted

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

Thank you for your interest in Nature Genetics.

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics
ORCiD: 0000-0003-1589-7087

Version 1:

Decision Letter:

28th Feb 2025

Dear Dr Sauler,

Your Article, "Aberrant Cellular Communities Define Disease Heterogeneity in Chronic Obstructive Pulmonary Disease" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

Briefly, the three reviewers seem positive for the value of your dataset and likely supportive of an eventual publication.

Reviewers #1 and #2 have primarily presentational requests; Referee #2 also asks about statistical rigour, in terms of the multiple tests performed, the test+validation experimental design, etc.
Reviewer #3 has more substantive comments, e.g. the comparison to healthy lung is needed, Xenium better-presented, and so forth.

In our reading all of these requests seem eminently addressable without a substantial expansion of the work or new data generation and we would, therefore, encourage you and your co-authors to respond as fully as possible.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

EXTENDED DATA FIGURES

When re-submitting your manuscript, please ensure that any supplementary figures and tables that are crucial to the manuscript's conclusions are converted into Extended Data figures and tables to increase visibility of these data. Extended Data figures and tables are online-only (present in the online PDF and full-text HTML versions of the paper), peer-reviewed display items that provide essential background to the article but are not included in the main article due to space constraints. A maximum of ten Extended Data display items (figures and tables) is permitted.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics
ORCiD: 0000-0003-1589-7087

Referee expertise: lung biology and disease, including -omics analysis thereof.

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

Zhang et al 2025 Nat Genet Manuscript Comments

The authors describe diverse and aberrant cell types underlying COPD at different stages of disease. While many previous studies have looked at COPD cells/ tissues, these have largely been done using specific omics techniques in isolation (e.g. scRNA-seq or spatial transcriptomics). The authors have integrated different omics (snRNA-seq, spatial transcriptomics, and proteomics to link specific cells types/ cell states to clinical features. The study is hugely significant and begins to address the complexity of COPD on a molecular and clinical level as well as providing an example of how to interrogate multi omics datasets to study complex diseases. It was particularly important that the authors linked specific findings to different clinical features, highlighting the complexity of, and the challenges of treating COPD.

The data has been obtained and analysed robustly and by integrating different types of omics data, the limitations of each method can be accounted for using other methods (i.e. applying spatial context to single cell data). I have a few comments on the GeoMx data analysis and the data QC, which are below.

My comments are mainly for clarity of the manuscript.

Comments:

One of my main comments is wondering why the authors chose to use snRNA-seq rather than scRNA-seq. This could be made clearer in the manuscript. Resources such as the human lung cell atlas generated from scRNA-seq has been used to identify different cell types in health and disease. Does the snRNA-seq data complement these different but similar datasets?

In the manuscript, authors combined never smokers with those who had no signs of airflow obstruction. Were the changes identified in COPD in reference to this group as a 'control' group. From table 1, it looks as if both these groups had no sign of airflow obstruction, but are there differences between the two groups that may impact the findings?

Authors have done lots of work to uncover complex cell types/ states/ interactions. Could they suggest any potential/ promising areas for future therapeutic intervention in the discussion i.e. how will these findings translate to clinic?

Can the authors elaborate on the marker genes used for the cell typing - is this determined from previously defined 'canonical' markers?

Clinical clustering – define which clinical characteristics are used.
Cluster 3 is people with severe disease but low symptom burden – how is 'severe disease' defined. It seems odd that someone would have severe disease but low symptom burden.

Is there differences in area of the lung that tissue was taken from. Would this likely affect cell composition and was this controlled for?

For the Xenium workflow, how was the 480 curated genes selected? Based on snRNA-seq data? Also for the spatial – were cores/ tissue sections selected based on tissue architecture/ histological features/ evidence of disease?

Line 350 - It would be beneficial to comment on the proteomics method performed in the results section as not clear. Which panel was used? Would one expect to see a big overlap with the transcriptomics data/ ECM proteomics given that proteomics panels are usually quite small (compared to whole transcriptome) - clarity on the proteins assayed would address this

Was plasma proteomics paired to the snRNA-seq and/or spatial donors? Can authors also define the OLink panel used – which proteins/ how many assayed.

Line 496 in the discussion authors mention that there is a discrepancy between AT2SCGB3A2 cells in advanced disease based on technique used (snRNA-seq or immunostaining) could this be a case of protein levels and mRNA levels not always correlating?

Line 575 – define which Novaseq machine (6000/ X Plus?)

For Xenium can authors expand on how the custom panel was chosen
For GeoMx, was the WTA used? Or was this custom/ smaller panel?
For GeoMx was there a criteria for drawing ROIs? E.g. minimum cell number/ area.
The analysis for the GeoMx – can the authors define which statistical test was used (I would guess using a linear mixed model). There are now packages available that seem to handle this type of data better than the standard GeoMx pipeline (PMID: 37953397).

Supplemental Figure 6 shows very high NTC counts for some of the libraries. Were these libraries used in the analysis, if so how was this controlled for? Obviously high contamination would cause considerable background noise. Could authors include images of representative ROIs for each region they were interested in. Could QC graphs also be separated to show QC for each region. In my experience some regions are 'more difficult' to capture than others.

Figure 1a) Clusters 3 and 4, and 6 and 7 look very similar – what were the differences? Could these be combined?

Figure 1b) I'm not sure this adds an awful lot to the overall figure and is probably unnecessary

Figure 2e) and follow up correlation plots – can the authors please clarify directionality of association. For Figure 2e for e.g. are more AT1 cells associated with higher DLCO? It isn't overly clear from the figures.

Figure 2i) I would suggest two UMAP plots for epithelial cells. One for alveolar and one for airway, rather than one for all and then one for airway specific.

Figure 2L) legend labelling and the figure should match

Figure 3a) this looks like a repeat of figure 2a? I don't think both are required

Figure 3 L-O) there is no O figure.

Figure 7 D-H) Re-define LI parenchyma and HI parenchyma here for clarity

Figure 8) I think the authors could make this figure clearer as to which were plasma proteins detected from the plasma proteomics, and which were ECM proteins.

Figure 9a) This is quite busy and hard to follow – I'm not sure it adds any value

Minor errors:
Repetition on line 570/571
Endothelial UMAPs the same – combine from figures 2 and 3 (2a and 2b – move to figure 3?)
Check text on figure 7 (clarity)
Line 227/228 – Provide acronym for aberrant basaloid cells (ABC)
Line 499 – grammar (cohort..)
Line 571 – sentence repeated
Line 738 – First sentence a bit repetitive (Xenium performed using Xenium)
Figure 7E) Check labelling
Figure 9) – figure E is missing from the legend.

Reviewer #2 (Remarks to the Author):

Zhang and colleagues report their findings from snRNA-seq of lung tissue samples from 141 subjects in the Lung Tissue Research Consortium, as well as spatial transcriptomics to confirm findings in an independent cohort of 37 subjects. The manuscript includes analyses investigating differences in expression profiles and cell states represented with respect to disease severity in COPD, information that will be of broad interest for the COPD research community. In general, I found the work to be of high quality, with reasonable methods applied in most cases, and interesting interpretation of the results.

Below, I provide some overall conceptual feedback for the authors regarding their general approach to reporting results (see point #1 below), as well as multiple detailed points for clarification.

Major:

1. It is commendable that the authors included in their study design an additional independent cohort of n=37 individuals for spatial transcriptomics in order to follow-up on their primary snRNA-seq effort. That being said, the Results of the manuscript overall read more descriptively, without a clear map of what were the specific hypotheses to be tested in the data presented. While I recognize there is a need for some descriptive presentation of the data resource presented in this manuscript, it would help to provide more details, regarding hypotheses to be tested and corresponding analytic approaches used to address those, where appropriate. In doing so, other issues may become apparent, including (a) the need for multiple comparisons corrections in statistical analyses, (b) the need for validation using independent data (including the n=37 from the separate cohort included in this study), and (c) the need for multiple comparisons corrections in examining the independent validation cohort. Currently, the manuscript glosses over these issues, which also leads to missed opportunities in terms of making certain points more concrete and precise.
2. Abstract and Methods: Please clarify whether all 141 subjects included in snRNA-seq were confirmed COPD cases, as the LTRC does include some participants with other disease states.
3. Methods: While the clustering (Fig 1A) provides information regarding comparison of the 141 LTRC subjects included in snRNA-seq as compared to the overall cohort, please additionally provide a description of how the participants were selected for sequencing in the current study.
4. Results / Methods: What was the motivation for having five subjects for whom two distinct lobes were sequenced? How did the downstream analysis account for the within individual correlation induced by this approach?
5. Methods: The Results section says "Key findings were further validated in an independent cohort of 37 subjects." Were these 37 subjects from the LTRC or a different cohort? If a different cohort, please provide methods regarding recruitment and study cohort description.
6. Table 1: The Results setion references a Table 1 of participants characteristics, but I could not find this Table in the submitted materials. Please submit the Table for peer review.
7. Table 1 and Supp Table 3: Please include a summary of the participant race/ancestry information for the overall snRNA-seq (n=141) and GeoMx Spatial Transcriptomics (n=37) cohorts.
8. Methods: The Results text says "Cell type annotation of the sequenced cells was made based on marker genes with stable expression across subjects" with a reference to Supp Table 1. While the list of marker genes is helpful, the following clarification is needed: (a) what criteria was used to identify marker genes with stable expression across subjects, and (b) what literature references or other approaches, if any, were used to assign labels to the identified clusters (e.g. ABC, AT1, AT2, Goblet, etc.)?
9. Results and Figures / Figure legends: Throughout the manuscript and in Figure legends, where p-values are reported, please state the name of the statistical test used to generate the p-value, as well as the sample size analyzed.
10. Figure 6C: Is it possible to add confidence bands to this line plot? Also, please specify units for the y-axis.
11. Results / Discussion: The text notes "ME12, enriched in antiprotease SERPINA1 and inhibitors … inversely correlated with inflamed non-immune cells" It is worth further contextualizing this finding either in the Results or Discussion, as SERPINA1 is the gene coding for Alpha-1 Antitrypsin, deficiency of which is the most well-documented genetic cause of

COPD and emphysema. In particular, are your findings consistent with the current understanding of AATD and its impact on risk of COPD / emphysema, or does your finding lead to new hypotheses in that regard?
12. Methods – Statistics: Please address the issue of multiple comparisons, including what statistical analyses in the manuscript require these considerations, and incorporate these corrections as appropriate in the Results and corresponding interpretation.

Minor:
1. Methods: There is a subsection labeled “Clustering of the entire LGRC dataset”. Please clarify what is the relationship between LTRC and LGRC.

Reviewer #3 (Remarks to the Author):

Summary of the key results:
Authors conducted snRNA-seq analysis of lung tissue from a large cohort of 141 subjects with COPD. They identified the shifts in cell composition and some pathologic cell states that correlates with clinical traits including lung function, emphysema, and symptoms. Some findings were validated through immunofluorescence and spatial transcriptomics. The study design is comprehensive, include proteomic profiling of the ECM and blood, and spatial transcriptomics and immunofluorescence of tissues to provided cross-multiomic validations and identify plasma biomarkers reflecting pathologic lung cell states. The study incorporates multiple analyses, including spatial niche analysis, mediation analyses and cell communication networks, to reveal signaling pathways driving disease. However, the present study raised several major concerns as discussed below.

Major comments:
1. Since the major impact of this manuscript is to map the COPD heterogeneity, to identify the shifts in cell composition and pathologic cell states that correlates with important clinical traits (i.e., lung function, emphysema, and symptoms, etc.), in this case, design to include normal lung control cohorts is necessary, the "shift" and "pathologic" states or “aberrant” cell populations can only be identified when control is present and compared. The integrated Human Lung Cell Atlas (HLCA, 37291214) or LungMAP CellRef (37516747) can be used for this purpose.

With the integration and comparison of the control lung data, authors might be able to identify pathologic cell types or states that are unique to COPD in addition to the aberrant basaloid cells (ABCs) and CTHRC1+ fibroblasts which were previously identified from pulmonary fibrosis studies. In addition, normal lungs can serve as important negative control to assess the inflammatory states in several cell types of COPD but not in the corresponding control cell types; and normal lung cell types or states do not correlate or correlate differently with clinical features.

2. Aberrant cellular communities are spatially organized within distinct neighborhoods, suggesting a critical role for cellular microenvironments in disease progression. This is an important finding of the paper. However, the figures from spatial transcriptome analysis are in very low resolution and cannot serve as strong supporting evidence. Xenium platform is capable to provide high resolution to detect marker expression at subcellular levels, together with quantitively measure of the cell selective markers expression and cell composition in representative niches (zoom in view) would be a better way to visualize the data.

3. Correlation of cell proportions with clinical features derived major findings in this work. However, a simple correlation analysis was used without considering the impact of covariates, such as age, sex, height, BMI, etc. on the analysis.

4. Authors revealed cellular pathways and signaling networks associated with clinical outcomes and inflammed cell states via connectome analysis. Again, it is important to show the comparison with normal controls. Does any of the ligand-receptor or signaling pathway prediction can be cross validated via proteomic profiling and spatial transcriptomics data?
5. Data available: data has not submitted to GEO or dbGAP.

Minor points:
1. Most of the figures were mixed with clear panels and fuzzy panels, some are unreadable (e.g., 1D, 3J, 4A, 4I, 7D-I, 9D, 9F and many supplementary panels)
2. Fig.1C: notes left in umap “change cluster number to final cell annotation”
3. The correlation between cell states and clinic severity measures are moderate, nothing above 0.5, which is commonly considered as the threshold for significant correlation.
4. Fig.3B, 3G: subtypes distinction is subtle. No selective markers to distinct them; not sure if these transitional states were only in COPD but not in control.
5. Fig 4I and J, the y axis values are confusing. They were labeled as “cell proportions (log2 scale)”. Proportion values should be within 0 and 1, and thus log2 scale will become negative values. However, in the y axis of Fig 4I and J, all log2 cell proportion values were positive.
6. "COL15A1+ systemic peribronchial endothelial cells increased with emphysema severity (P=0.003) (Fig. 2E)." systemic peribronchial endothelial cells were increased with emphysema stage in Fig 2b & 2C, not Fig.2E
7. EXT Fig 9 shows the umaps of Xenium by lineages, can you make a complete umap with marker genes support?

Version 2:

Decision Letter:

Our ref: NG-A67458R1

22nd Sep 2025

Dear Dr Sauler,

Thank you for submitting your revised manuscript "Aberrant Cellular Communities Define Disease Heterogeneity in Chronic Obstructive Pulmonary Disease" (NG-A67458R1). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Safia Danovi, PhD
Senior Editor, Nature Genetics
ORCiD: 0009-0007-7822-5479

Reviewer #1 (Remarks to the Author):

The authors have made substantial changes to the manuscript, which satisfactorily address my comments made during my review. The new manuscript has improved readability and clarity, particularly in the methods used.

Reviewer #2 (Remarks to the Author):

The authors should be commended for their thorough responses to the comments from prior review. I do not have any additional comments at this time.

Reviewer #3 (Remarks to the Author):

This study provides a comprehensive, high-resolution cell atlas of COPD by integrating snRNA-seq of 141 patients with spatial transcriptomics, proteomics, and extensive clinical data. While the overwhelming complexity of the design and interpretation still raised some concern, overall, I am happy with the revised version and think the paper will be useful to advance our understanding of COPD.

Response to the Reviewers' Comments for "Aberrant Cellular Communities Define Disease Heterogeneity in Chronic Obstructive Pulmonary Disease"

General Comments

Briefly, the three reviewers seem positive for the value of your dataset and likely supportive of an eventual publication.

Reviewers #1 and #2 have primarily presentational requests; ***Referee #2*** also asks about statistical rigour, in terms of the multiple tests performed, the test+validation experimental design, etc.

Reviewer #3 has more substantive comments, e.g. the comparison to healthy lung is needed, Xenium better presented, and so forth.

In our reading all of these requests seem eminently addressable without a substantial expansion of the work or new data generation, and we would, therefore, encourage you and your co-authors to respond as fully as possible.

We appreciate the overall positive comments as well as the constructive feedback which we feel strengthened the clarity of the manuscript. Below is a point-by-point response.

Reviewer 1:

The authors describe diverse and aberrant cell types underlying COPD at different stages of disease. While many previous studies have looked at COPD cells/tissues, these have largely been done using specific omics techniques in isolation (e.g. scRNA-seq or spatial transcriptomics). The authors have integrated different omics (snRNAseq, spatial transcriptomics, and proteomics to link specific cell types/cell states to clinical features. The study is hugely significant and begins to address the complexity of COPD on a molecular and clinical level as well as providing an example of how to interrogate multi omics datasets to study complex diseases. It was particularly important that the authors linked specific findings to different clinical features, highlighting the complexity of, and the challenges of treating COPD.

The data has been obtained and analysed robustly and by integrating different types of omics data, the limitations of each method can be accounted for using other methods (i.e. applying spatial context to single cell data). I have a few comments on the GeoMx data analysis and the data QC, which are below.

My comments are mainly for clarity of the manuscript.

Comments:

Comment 1: One of my main comments is wondering why the authors chose to use snRNAseq rather than scRNA-seq. This could be made clearer in the manuscript. Resources such as the human lung cell atlas generated from scRNA-seq has been used to identify different cell types in health and disease. Does the snRNAseq data complement these different but similar datasets?

Response 1:

We chose snRNAseq for three practical reasons. First, and most importantly, it does not require fresh tissue and can be applied to archived frozen samples, which allowed us to analyze the extensively phenotyped LTRC biobank. Second, it reduces dissociation bias and captures fragile epithelial populations that are easily lost with whole-cell methods, providing a more representative view of lung tissue. Third, direct comparisons show that snRNAseq detects as many unique genes as matched scRNA-seq in lung tissue (Koenitzer et al., Am J Respir Cell Mol Biol 2020, PMID 32804550; reference 43 in our manuscript).

In response to the reviewer's comments, we have elaborated on this choice in the Introduction ("Isolating nuclei for sequencing enabled profiling of archived lung tissue from a large, well-phenotyped cohort…") and in the Discussion.

Public resources such as the Human Lung Cell Atlas have been invaluable for cell type annotation and served as essential references for validating our annotations. In many cases, our canonical marker genes showed strong concordance with those reported in these datasets, lending further confidence to the accuracy of our assignments. We have now commented on this in the Methods and Results sections, stating: "Clusters were annotated by automated label transfer and fine-tuned based on marker genes referenced from the Human Lung Cell Atlas (HLCA) and LungMAP."

However, most large-scale lung cell atlases to date have primarily focused on healthy lungs, with limited representation of COPD cases. Even when COPD subjects have been included, the extent of phenotypic characterization is often minimal, typically lacking standardized physiologic data such as pulmonary function tests (PFTs), imaging (e.g., CT-defined emphysema), and symptom measures. This is a critical limitation, as COPD is not a binary trait but a heterogeneous continuum of pathologic and clinical features that cannot be fully captured by a simple "COPD vs. non-COPD" label.

Our previous scRNA-seq study (Sauler et al., Nat Comm, 2022) illustrates this limitation—although it was among the first to focus specifically on COPD tissue, it included only a small number of end-stage lungs and lacked complete physiologic or radiologic profiling. In contrast, the current study analyzes 1.52 million nuclei from 141 deeply phenotyped individuals across GOLD stages 0–IV, integrating structural and immune compartments. This comprehensive approach allows us to detect disease-associated cell states, including both inflammatory and reparative populations, not previously identified as COPD-relevant in prior references.

Taken together, our findings complement and extend existing resources by anchoring new disease-associated states within well-characterized clinical and pathologic contexts. In addition, we have deposited these data into public repositories (including the HLCA) and the NIH's Sequence Read Archive (SRA) database.

Comment 2: In the manuscript, authors combined never smokers with those who had no signs of airflow obstruction. Were the changes identified in COPD in reference to this group as a 'control' group. From table 1, it looks as if both these groups had no sign of airflow obstruction, but are there differences between the two groups that may impact the findings?

Response 2:

We appreciate the opportunity to clarify our analytic strategy, which was designed to interrogate COPD heterogeneity from multiple complementary angles.

We initially requested never-smokers, smokers without disease, and those with increasing GOLD stage from the NHLBI BioLINCC repository, which is why we structured our table as we did.

First, we approached COPD along individual clinical domains, including GOLD stage, emphysema severity, smoking history, and symptom burden, using univariate analyses to identify cell populations associated with specific features of disease (Figures 2–3). Indeed, we found features of disease that were independently associated with specific aspects of pathology. For example, goblet cell abundance correlated most strongly with smoking history, whereas endothelial populations were more tightly linked to radiographic emphysema. These analyses allowed us to map cell-state associations to distinct clinical axes—a useful starting point for deconstructing the complexity of COPD. We did not identify novel shifts in cell populations between never-smokers and "healthy" smokers beyond those identified when assessing GOLD criteria and smoking independently.

To address the reviewer's point, we also examined composite clinical traits. We performed an unsupervised clustering analysis of clinical traits across the entire 2,200-patient LTRC cohort, aiming to define stable clinical phenotypes that could serve as an external reference for the 141-patient snRNAseq cohort. We did this for several reasons:

1. The clinical traits used to assess COPD are not independent; they are often collinear, both in our cohort and in larger studies. While trait-specific analyses are informative, they fail to capture how combinations

of features define clinically meaningful subgroups. By leveraging clinical clustering, we sought a more holistic framework to contextualize cell-state variation within the multifaceted nature of COPD.

2. We aimed to increase the generalizability of our findings by anchoring the single-cell dataset to a broader, clinically diverse population. Given the modest size of the 141-patient snRNAseq cohort, we recognized the possibility of sampling bias. Clinical clustering based on the full 2,200-patient cohort allowed us to more confidently infer how cell states might map onto widely observed patient phenotypes.

This approach yielded meaningful insights. For example, we identified inflammatory cell populations enriched in a clinical subtype characterized by severe physiologic impairment and high symptom burden—distinct from another group with similar physiologic deficits but low symptom scores. This distinction has potential therapeutic implications, as it highlights the cellular correlates of symptom severity beyond airflow limitation alone. We have described this in more detail in the Online Methods and have amended the Results to reflect these two goals.

Comment 3: Authors have done lots of work to uncover complex cell types/states/interactions. Could they suggest any potential/promising areas for future therapeutic intervention in the discussion i.e. how will these findings translate to clinic?

Response 3: We have revised the Introduction and Discussion to more clearly highlight the translational implications of our findings and to emphasize how our cellular phenotyping approach advances precision medicine for COPD. We believe it is most boldly stated in the concluding paragraph: *"… this work offers a framework for understanding COPD heterogeneity rooted in fundamental lung biology, extending beyond traditional physiologic or symptom-based classifications. By linking specific aberrant cell populations to clinical features, we highlight pathways with potential for therapeutic targeting. The association of a plasma protein module with inflamed endothelial cells provides proof-of-concept that systemic biomarkers can reflect localized inflammatory niches, supporting the feasibility of longitudinal, non-invasive disease assessment. For clinicians, these findings offer a roadmap toward mechanism-guided patient stratification, aligning the right treatment to the predominant biological processes driving disease in each individual, and ultimately advancing COPD care toward a precision medicine model."*

Comment 4: Can the authors elaborate on the marker genes used for the cell typing - is this determined from previously defined 'canonical' markers?

Response 4: We have improved the clarity of the marker gene section in our Methods and Results section. In the results section we write. Clusters were annotated by automated label transfer and fine-tuned based on marker genes referenced from the human lung cell atlas (HLCA) and LungMAP. In the Methods we write Cell identities were assigned by mapping to the Human Lung Cell Atlas using Azimuth and defining clusters by the prevailing Azimuth (v0.5.0) annotation complemented by manual annotation based on commonly cited genes [51]. To ensure clustering was not influenced by technical covariates, all reported clusters were visually inspected for differences in gene count, feature count, percent mitochondrial fraction, and antisense genes.

Comment 5: Clinical clustering – define which clinical characteristics are used.

Response 5: We apologize for the lack of clarity. Previously this analysis was split into two locations. Now, we restructure this analysis putting it all in one location. In the results section for Figure 4, we now write**:**

*While single-trait analyses identified associations, they could not address how cell populations relate to composite COPD phenotypes. We therefore performed unsupervised clustering of 110 categorical and 24 continuous variables from the full 2,213-subject LTRC cohort, including spirometry, radiographic emphysema, smoking history, symptoms, and functional capacity (**Supplemental Table 5**), and projected these phenotypes onto our 141-subject snRNA-seq dataset. This framework identified multi-trait disease subgroups within a larger, more representative population, supporting broader applicability and helping to mitigate sampling bias.*

In the Methods section we write "

Clinical trait-based clustering of the LTRC dataset. *Clustering of variables was performed using the ClustOfVar [43] R package. We retained 110 categorical and 24 continuous variables that were available across all the patients. Physiologic and exposure features, treated as continuous variables, included FEV1/FVC ratio,*

FEV1 percent predicted, FVC percent predicted, diffusing capacity (DLCO percent predicted), smoking metrics (cigarettes per day, cumulative pack-years, years since cessation), and age. Radiographic structure was summarized with the LTRC semi-quantitative CT descriptions for features that include emphysema, airway inflammation, mosaic attenuation, nodules, ground glass, etc. Patient-reported outcomes were captured with the St. George's Respiratory Questionnaire (symptoms, activity, impacts, and total score), the SF-12 from which we used eight norm-based domain scores and the Physical (PCS) and Mental (MCS) component summaries; questionnaire items in Supplementary Fig. 2 that contributed to the domains: Dyspnea, Cough, Infection, Exacerbation, Wheezing. The full list of LTRC variables used for clustering is listed in ***Supplemental Table 5"***

Comment 5 (cont): Cluster 3 is people with severe disease but low symptom burden – how is 'severe disease' defined. It seems odd that someone would have severe disease but low symptom burden.

Response 5 (cont): Cluster 3 was characterized by subjects with severe physiologic abnormalities, such as reduced FEV_1 and extensive radiographic emphysema—but relatively low reported symptom burden. While this may seem paradoxical, it highlights a well-recognized aspect of COPD heterogeneity. A dissociation between objective physiologic impairment and subjective symptomatology is acknowledged in clinical practice and reflected in major guidelines such as GOLD, which recommends stratifying patients along both axes to guide treatment. Consequently, patients can have similar degrees of spirometric impairment and radiographic emphysema, but one subject may have more subjective dyspnea, exacerbations, etc. Our clustering recapitulates this disconnect, underscoring its biological and clinical relevance and supporting the need for multidimensional approaches to COPD classification. This approach yielded meaningful insights. For example, we identified inflammatory cell populations enriched in a clinical subtype characterized by severe physiologic impairment and high symptom burden, which was distinct from the group with similar physiologic deficits but low symptom scores. This distinction has potential therapeutic implications, as it highlights the cellular correlates of symptom severity beyond airflow limitation alone. We have now adjusted the names of these clusters in the text to distinguish physiologic from symptomatic severity**.**

Comment 6: Are there differences in area of the lung that tissue was taken from. Would this likely affect cell composition and was this controlled for?

Response 6: Regarding tissue sampling, non-diseased and early-stage disease samples were more commonly obtained from subjects undergoing lobectomies, presumably for peripheral nodules, while late-stage disease samples were more commonly obtained from either lung explants or lung volume reduction surgeries. All tissue samples were collected from regions at least 10 cm away from any tumor or nodule. We have now included the lobes from which the samples were obtained in all Tables describing the clinical characteristics of the subjects.

We recognize regional differences in lung structure and disease distribution, particularly the upper-lobe predominance of emphysema, may also influence cellular composition. To partially account for this, we restricted our emphysema analysis to lobe-specific radiographic data. Additionally, for analyses performed at the level of the lobe (radiographic emphysema), we now adjust for the anatomic lobe.

Comment 7: For the Xenium workflow, how was the 480 curated genes selected? Based on snRNAseq data? Also, for the spatial – were cores/tissue sections selected based on tissue architecture/histological features/evidence of disease?

Response 7: The custom gene panel for Xenium was developed in collaboration with 10x Genomics and was informed by our snRNAseq dataset. Specifically, we prioritized genes that served three key purposes: (1) robustly mark canonical lung cell types across epithelial, endothelial, fibroblast, and immune compartments; (2) enable detection and spatial mapping of disease-associated cell states identified in our study, including aberrant basaloid cells (ABCs), $CTHRC1^+$ fibroblasts, and inflamed non-immune populations; and (3) include key signaling molecules involved in cell-cell and cell-matrix interactions, based on pathway analyses.

We have amended the results and the Methods to reflect this.

For spatial profiling, tissue cores were selected to ensure representative sampling of both parenchymal regions and structures such as airways and/or large vessels. All samples underwent histopathologic review by a pulmonary pathologist to assess relevance for downstream spatial analyses. We have amended the Methods to reflect this.

Comment 8: Line 350 - It would be beneficial to comment on the proteomics method performed in the results section as not clear. Which panel was used?

Would one expect to see a big overlap with the transcriptomics data/ECM proteomics given that proteomics panels are usually quite small (compared to whole transcriptome) - clarity on the proteins assayed would address this.

Response 8: We thank the Reviewer for this comment. Proteomic profiling of the lung was performed on decellularized lung ECM samples using LC-MS/MS. In contrast, plasma proteomics was performed using the Olink Explore HT panel (see **Supplemental Table 7A** for the full list). We have revised the methods and results section for clarity. And make explicit mention of using the Olink Explore HT panel on matched plasma samples

The reviewer is correct in noting the difference in scale between the transcriptomic and proteomic platforms. While single-cell and single-nucleus RNA sequencing detect over 30,000 genes, the Olink Explore HT panel measures 5,420 protein biomarkers. Beyond this inherent limitation, proteomic analyses were necessarily performed at the subject level rather than the cellular level, and our sample size of 63 plasma profiles limited statistical power to detect associations with individual proteins. We make this more clear in the text and write "*To assess whether aberrant cell states were reflected in circulation, we profiled matched plasma from 64 snRNA-seq subjects using the Olink Explore HT panel (5,420 protein biomarkers) and applied a module-based approach to account for the smaller protein panel relative to the transcriptome and the limited sample size."*

Comment 9: Was plasma proteomics paired to the snRNAseq and/or spatial donors? Can authors also define the OLink panel used – which proteins/ how many assayed.

Response 9: Plasma proteomics was performed on 64 plasma samples paired with the snRNAseq dataset. There was significant overlap between the plasma proteomics and spatial transcriptomics, however we did not evaluate this in our manuscript. We make this more abundantly clear in the third paragraph of the results section. We now mention the Olink Explorer HT panel in our dataset.

Comment 10: Line 496 in the discussion authors mention that there is a discrepancy between AT2SCGB3A2 cells in advanced disease based on technique used (snRNAseq or immunostaining) could this be a case of protein levels and mRNA levels not always correlating?

Response 10: We thank the Reviewer for this observation. We agree that the discrepancy between our findings and prior studies could, in part, be due to differences at the gene and protein level, which may not always correlate. We have added to the discussion to specify this point. Specifically, we state "*While our data and prior reports show AT2$_{SCGB3A2}$ cells increase in COPD, we observed a modest decline in advanced disease, differing from prior reports.[11,12] This difference may have arisen from methodological differences, such as the narrow transcriptional window captured by snRNA-seq compared to protein detection by immunostaining, or differences in cohort composition, smoking status, or disease severity.*

Comment 11: Line 575 – define which Novaseq machine (6000/ X Plus?)

Response 11: We have updated the methods section to specify sequencing was performed using the NovaSeq 6000 system.

Comment 12: For GeoMx, was the WTA used? Or was this custom/ smaller panel? For GeoMx was there a criteria for drawing ROIs? E.g. minimum cell number/ area.The analysis for the GeoMx – can the authors defines which statistical test was used (I would guess using a linear mixed model). There are now packages available that seem to handle this type of data better than the standard GeoMx pipeline (PMID: 37953397).

Response 12: The reviewer is correct and the WTA was used for the GeoMx. We also added a description of QC criteria in the Method section including nuclei counts and area. Specifically, segments with < 100 nuclei or an area < 5,000 μm² were excluded.

To perform deconvolution, we used spatialDWLS, which extends the dampened weighted least squares (DWLS) framework to spatial data. After getting estimated cell type proportions, group-wise differences in inflammatory cell proportions were first assessed using the Kruskal-Wallis test. Post-hoc pairwise comparisons were

performed using Wilcoxon rank-sum tests with FDR correction. This is now described in every figure legend in which we present this data.

We thank the reviewer for pointing out that new tools are now available to analyze GeoMx data. In this case, since the dataset is publicly available and already pre-processed, we followed the original authors' pipeline to maintain consistency with the published results.

Comment 13: Supplemental Figure 6 shows very high NTC counts for some of the libraries. Were these libraries used in the analysis, if so, how was this controlled for? Obviously high contamination would cause considerable background noise.Could authors include images of representative ROIs for each region they were interested in. Could QC graphs also be separated to show QC for each region. In my experience some regions are 'more difficult' to capture than others.

Response 13: The reviewer is correct that some of those NTC counts are indeed very high. NTC counts exceeding 1,000 were excluded from downstream analysis, and we apologize if that was not clear during the initial submission. Following your suggestion, we generated QC figures for each tissue region (**Supplementary Fig. 6**). We expanded our Methods section as well. We write *We applied continuous QC filters to all 924 ROIs at the segment (AOI) level prior to analysis. ROIs were excluded if they failed any of the following criteria: raw reads <1,000; trimming rate <80%; stitching rate <80%; alignment rate <80%; sequencing saturation <50% (computed as 1−deduplicated reads/aligned readsx100); negative-control geometric mean (non-targeting probes) <1; no-template control (NTC) count >1,000; nuclei count <100; segmented area <5,000 μm², and <1% genes detected…A total of 467 ROIs from four tissue locations (airway, follicle, parenchyma, and vessel) passed quality control. were included in downstream analyses. Notably, this led to the exclusion of samples (33 samples were retained*). Representative ROIs were not available for this public dataset (GSE237120). We also summarized the proportion of segments removed during filtering. We write… "*The deletion rates, from highest to lowest: vessel (61%), airway (55%), follicle (47%), and parenchyma (32%).*" This is consistent with your observation that some regions are more difficult to capture. This has been added to the methods.

Comment 14: Figure 1a) Clusters 3 and 4, and 6 and 7 look very similar – what were the differences? Could there be combined? Figure 1b) I'm not sure this adds an awful lot to the overall figure and is probably unnecessary

Response 14: We appreciate the reviewer's observation regarding the apparent similarity between clusters 3 and 4, and clusters 6 and 7 in Fig. 1A. These clusters were initially retained as separate groups because they represented the results of the clustering, and so we wanted to report that analysis as accurately as possible. However, we agree that the distinctions are subtle and, upon further evaluation, concluded that combining these clusters would improve clarity and interpretability without compromising biological or clinical resolution. Indeed, when we assess the cluster assignment for our sequenced samples, we combined Clusters 3 and 4 together, and Clusters 6 and 7 together. A collective description of these results is now provided in the Results section discussing **Figure 4A-D**.

We have also removed **Fig. 1B** from the figure since we agree that it did not substantially contribute to the overall message of **Fig. 1**.

Comment 15: Figure 2e) and follow up correlation plots – can the authors please clarify directionality of association. For Figure 2e for e.g. are more AT1 cells associated with higher DLCO? It isn't overly clear from the figures.

Response 15: We have now improved clarity in the figure legend of the manuscript. We meant to imply that the proportion of AT1 cells decreases with increasing GOLD stage and a decline in DLCO, suggesting that the loss of AT1 cells is associated with disease progression. We recognize that the original presentation may have been unclear, we have revised the text throughout the manuscript for improved clarity.

Figure 2i) I would suggest two UMAP plots for epithelial cells. One for alveolar and one for airway, rather than one for all and then one for airway specific.

We agree with the reviewer that this would make more sense. However, it obscures the relationship between AT2 SCGB3A2^{+} cells, AT2 cells, and airway cells a connection we believe is important given their biology and their

initial discovery in the context of COPD. Additionally, the UMAP of AT2 and AT1 cells does not exhibit clearly delineated regions but rather each cell is a continuum of transcriptional features – limiting the impact of showing the UMAP. For these reasons, we elected to maintain the current labeling strategy, but we include a UMAP of all the epithelial cells in which AT2 and AT1 cells can be readily discerned.

Figure 2L) legend labelling and the figure should match

We have now corrected this.

Comment 16: Figure 3a) this looks like a repeat of figure 2a? I don't think both are required. , Figure 3 L-O) there is no O figure. :

Response 16: Thank you for pointing these out. We have now merged Figure 2a and 3a into figure 1, delineating inflamed cells within their respective lineages. We have corrected the Figure O omission.

Comment 17: Figure 7 D-H) Re-define LI parenchyma and HI parenchyma here for clarity

Response 17: We have ensured that low immune (LI) and high immune (HI) parenchyma are defined in the figure caption.

Comment 18: Figure 8) I think the authors could make this figure clearer as to which were plasma proteins detected from the plasma proteomics, and which were ECM proteins.

Response 18: We thank the Reviewer for their suggestion. We have revised the panel and caption to more clearly distinguish between ECM and plasma proteins for improved clarity.

Comment 19: Figure 9a) This is quite busy and hard to follow – I'm not sure it adds any value

Response 19: We appreciate the Reviewer's feedback and recognize that **Fig. 9A** is quite busy and could be difficult to follow. To address this, we have removed the figure from the **Fig. 9** panel and placed in the Extended Data section.

Minor Errors:

Comment 20:

1) Repetition on line 570/571
2) Endothelial UMAPs the same – combine from figures 2 and 3 (2a and 2b – move to figure 3?)
3) Check text on figure 7 (clarity)
4) Line 227/228 – Provide acronym for aberrant basaloid cells (ABC)
5) Line 499 – grammar (cohort.)
6) Line 571 – sentence repeated
7) Line 738 – First sentence a bit repetitive (Xenium performed using Xenium)
8) Figure 7E) Check labelling
9) Figure 9) – figure E is missing from the legend.

Response 20: We thank the Reviewer for the detailed and helpful comments. Please find our point-by-point responses below:

1) We have revised the text to eliminate any potential redundancy and to improve clarity.
2) This has been addressed in our response to *Reviewer 1 – Comment #11*.
3) Thank you for bringing this to our attention, we have reviewed and revised **Fig. 7** text for clarity and accuracy.
4) We have defined the acronym (ABC) upon the first mention of aberrant basaloid cells.
5) This grammatical issue has been corrected.
6) We have revised the text to eliminate any redundancy and improve flow.
7) The sentence has been revised for clarity and to remove repetition.
8) The labeling has been reviewed and corrected.
9) Thank you for catching this oversight. The legend has been updated to include a description for panel E.

Reviewer 2:

Zhang and colleagues report their findings from snRNAseq of lung tissue samples from 141 subjects in the Lung Tissue Research Consortium, as well as spatial transcriptomics to confirm findings in an independent cohort of 37 subjects. The manuscript includes analyses investigating differences in expression profiles and cell states represented with respect to disease severity in COPD, information that will be of broad interest for the COPD research community. In general, I found the work to be of high quality, with reasonable methods applied in most cases, and interesting interpretation of the results.

Below, I provide some overall conceptual feedback for the authors regarding their general approach to reporting results (see point #1 below), as well as multiple detailed points for clarification.

Major:

Comment 1: It is commendable that the authors included in their study design an additional independent cohort of n=37 individuals for spatial transcriptomics in order to follow-up on their primary snRNAseq effort. That being said, the Results of the manuscript overall read more descriptively, without a clear map of what were the specific hypotheses to be tested in the data presented. While I recognize there is a need for some descriptive presentation of the data resource presented in this manuscript, it would help to provide more details, regarding hypotheses to be tested and corresponding analytic approaches used to address those, where appropriate.

Response 1: The concern raised by the reviewer has been the subject of considerable internal discussion, particularly regarding how best to present a complex, multiomic dataset in a manner that adheres to the scientific method while remaining readable and accessible.

Our methodological approach to clustering and sub-clustering was guided by biologically grounded hypotheses. Specifically, we hypothesized that distinct, disease-associated cell states would emerge within canonical lineages, and that these states would vary in abundance across the clinical spectrum of COPD. To test this, we applied lineage-specific clustering, allowing us to identify both known and novel cell states. At each step, we assessed whether these populations were enriched in disease, how they correlated with clinical traits, and whether their relevance could be validated using orthogonal approaches such as spatial transcriptomics and proteomics. Thus, clustering was not a purely data-driven exercise, but rather a hypothesis-guided framework rooted in prior knowledge of lineage identity, cell function, and COPD pathobiology.

Initially, we structured the Results section to follow this logic explicitly—for example, characterizing endothelial cells, identifying disease-enriched subsets, assessing correlations, and validating findings. While this approach mirrored the scientific method, the repetitive format across multiple cell lineages became difficult to follow. For instance, describing each inflamed non-immune cell across every lineage or reporting the identification of ABCs and CTHRC1^{+} fibroblasts per lineage introduced redundancy, rather than integrating these findings in the context of their known co-association.

To improve readability, we previously reorganized the Results to more efficiently synthesize multiple lines of evidence. We recognize, however, that this structure may have obscured the hypothesis-driven nature of our analysis. In response to the reviewer's feedback, we have revised the Results to more clearly articulate the hypotheses underpinning each major section. We now explicitly state the rationale for clustering strategies, the questions addressed, and the analytical approaches used. We believe this revision restores scientific clarity while preserving a coherent and accessible narrative, and we hope the revised structure meets the reviewer's expectations.

In doing so, other issues may become apparent, including (a) the need for multiple comparisons corrections in statistical analyses, (b) the need for validation using independent data (including the n=37 from the separate cohort included in this study), and (c) the need for multiple comparisons corrections in examining the independent validation cohort. Currently, the manuscript glosses over these issues, which also leads to missed opportunities in terms of making certain points more concrete and precise.

We thank the reviewer for this important point. We have improved the clarity of our analytic structure in the Methods (described above). Importantly, we have carefully reanalyzed all statistical associations presented in the manuscript and have updated the text and figure legends to clarify that all p-values have been adjusted for multiple comparisons where appropriate, using the Benjamini-Hochberg method. Importantly, we identified a method to analyse compositional data while adjusting for co-variates and performed these tests with input from

the author of the manuscript where this technique is first described (sccomp method). These adjustments have been consistently applied to all compositional analyses, correlation tests, and pathway enrichments described in the primary dataset. We believe these changes address the reviewer's concerns and improve the clarity, transparency, and interpretability of our results.

Comment 2: Abstract and Methods: Please clarify whether all 141 subjects included in snRNAseq were confirmed COPD cases, as the LTRC does include some participants with other disease states.

Response 2: We have now placed the table of the clinical characteristics of the subjects directly into the main text and below for clarity. This includes 13 never smokers and 15 individuals with normal spirometry.

Subject Characteristics															
Group	n	FEV1 % predicted [IQR]	FEV1/ FVC [IQR]	DLCO % predicted [IQR]	Emphysema %HU<−950 [IQR]	Pack years [IQR]	Quit years [IQR]	BMI [IQR]	Age [IQR]	SGRQ Symptoms [IQR]	Collection method	Self-identified race	Smoking status	Sex	Lobe
Never smoker	13	87 [16]	0.76 [0.14]	0.75 [0.07]	0.04 [0.04] (8)	0 [0]	NA [NA] (13)	30.65 [8.95]	58 [26]	6 [17]	L: 13	AA: 2; W: 11	N: 13; F: 0; C: 0	F: 9; M: 4	LG: 1; LL: 1; LU: 4; RL: 3; RM: 1; RU: 3
Normal spirometry	15	96 [9.5]	0.74 [0.06]	0.64 [0.18]	0.06 [0.04] (7)	40 [40.38]	13.42 [17.42]	28.61 [5.65]	67 [13]	11 [18.5]	L: 15	AA: 1; W: 14	N: 0; F: 14; C: 1	F: 7; M: 8	LL: 4; LU: 3; RL: 1; RM: 2; RU: 5
GOLD I, II	29	67 [20]	0.59 [0.12]	0.56 [0.22] (1)	0.11 [0.14] (7)	45 [35]	1.71 [10.49]	26.07 [7.23]	65 [12]	23 [37]	E: 4; L: 24; LVRS: 1	AA: 1; W: 28	N: 0; F: 22; C: 7	F: 14; M: 15	LL: 4; LU: 16; RL: 2; RM: 2; RU: 8
GOLD III	25	36 [9]	0.4 [0.13]	0.31 [0.17] (1)	0.3 [0.24] (8)	36 [20]	5.94 [8.64]	27.34 [8.8]	68 [8]	61 [34]	E: 8; L: 9; LVRS: 8	AA: 5; W: 20	N: 0; F: 22; C: 3	F: 15; M: 10	LL: 3; LU: 9; RL: 3; RM: 3; RU: 7
GOLD IV	59	20 [7.67]	0.28 [0.09]	0.25 [0.12] (9)	0.39 [0.18] (20)	40 [35]	5.45 [7.87]	24.85 [6.32]	60 [7.5]	53 [31]	E: 42; L: 1; LVRS: 16	AA: 3; M: 1; W: 55	N: 0; F: 59; C: 0	F: 31; M: 28	LG: 2; LL: 11; LU: 18; RL: 6; RM: 9; RU: 15

Table 1: Characteristics of samples obtained from subjects undergoing single-nucleus RNA sequencing. Results reflect median and interquartile range (IQR) for continuous variables or the number of subjects per group for categorical variables. () = missing variables. GOLD: Global Initiative for Chronic Obstructive Lung Disease criteria for disease severity. Smoking status: N = never, F = former, C = current; Sex: M = male, F = female, Collection method: L = lobectomy, E = explant, LVRS = lung volume reduction surgery; Self-identified race: AA = Black or African American, W = White, M = Multiracial; SGRQ = St. George's Respiratory Questionnaire; HU = Hounsfield units.

Comment 3: Methods: While the clustering (Fig 1A) provides information regarding comparison of the 141 LTRC subjects included in snRNAseq as compared to the overall cohort, please additionally provide a description of how the participants were selected for sequencing in the current study.

Response 3: To ensure representation across the full clinical spectrum of COPD, we submitted study requests to the LTRC biorepository specifically asking for lung tissue samples from subjects spanning all GOLD stages (0–IV). The final set of 141 subjects selected for snRNAseq was determined by the availability and quality of tissue samples meeting these criteria. This sampling strategy enabled downstream analyses to interrogate cell states in relation to disease progression. We have now included this update in the Methods section.

Comment 4: Results/Methods: What was the motivation for having five subjects for whom two distinct lobes were sequenced? How did the downstream analysis account for the within individual correlation induced by this approach?

Response 4: The motivation for sequencing two distinct lobes from five subjects was to assess stability of compositional changes across lobes. We now address the answer to this question in the Discussion: "

When we performed snRNAseq analysis of multiple samples derived from the same lobe, we observed consistent compositional patterns, whereas comparisons across different lobes suggested compositional differences, although our current dataset was underpowered to conclusively demonstrate this. Nevertheless, the identification of significant correlations between plasma inflammatory biomarkers and inflamed endothelial cell populations suggests the presence of a dominant pathway, and future studies with enhanced statistical power will be required to clarify these associations."

For downstream analyses linking samples to systemic clinical traits, we employed a per-subject approach to avoid bias introduced by within-subject correlations. However, for emphysema-focused analyses, in which the lobe-specific features were analyzed, we adopted a per-sample approach, as individual lobes exhibited varying degrees of disease involvement, and these samples represented only a small subset of the overall dataset. We now state this explicitly in both the Methods and the Results section.

Comment 5: Methods: The Results section says, “Key findings were further validated in an independent cohort of 37 subjects.” Were these 37 subjects from the LTRC or a different cohort? If a different cohort, please provide methods regarding recruitment and study cohort description.

Response 5: The subjects were part of an independent cohort. The study population is outlined in the manuscript Rojas-Quintero et al, Spatial Transcriptomics Resolve an Emphysema-Specific Lymphoid Follicle B Cell Signature in Chronic Obstructive Pulmonary Disease”, AJRCCM 2024. We have made this clearer in the text (describe it as the Baylor cohort) and provided more detail from this manuscript in the Methods regarding the recruitment and study cohort description. We have also included a table reflecting the clinical characteristics of the patients (**Supplemental Table 2**). Finally, in the process of revising, we realized that we reported the number of subjects incorrectly (as we included subjects in the final count that were excluded from analysis after QC). We have corrected this error and report the number of subjects as 33.

Comment 6: Table 1: The Results section references a Table 1 of participants characteristics, but I could not find this Table in the submitted materials. Please submit the Table for peer review.

Response 6: We apologize for this oversight. **Table 1** is now embedded within the body of the text.

Comment 7: Table 1 and Supp Table 3: Please include a summary of the participant race/ancestry information for the overall snRNAseq and GeoMx Spatial Transcriptomics cohorts.

Response 7: We have now updated all tables to include self-reported race for subjects in both the snRNAseq and GeoMx Spatial Transcriptomics cohorts.

Comment 8: Methods: The Results text says “Cell type annotation of the sequenced cells was made based on marker genes with stable expression across subjects” with a reference to Supp Table 1. While the list of marker genes is helpful, the following clarification is needed:

(a) what criteria was used to identify marker genes with stable expression across subjects, and (b) what literature references or other approaches, if any, were used to assign labels to the identified clusters (e.g. ABC, AT1, AT2, Goblet, etc.)?

Response 8:

We have now improved the clarity of our description of marker gene selection. Marker genes were identified through differential expression analysis within each major lineage to determine cluster-specific markers. Genes were prioritized based on their uniqueness to a given cluster, magnitude of fold change compared to other clusters, and statistical significance of enrichment. We also qualitatively ensured that marker genes were not primarily driven by individual subjects, confirmed by diffuse expression patterns across multiple subjects in both UMAP visualizations and heatmaps grouped by subject. Given that we did not explicitly quantify inter-subject variability or establish precise thresholds for expression stability, we have revised the text to remove reference to stable expression.

Importantly, the marker genes selected for canonical cell populations align with previously published literature, including reference atlases of human lung single-cell transcriptomes such as the Human Lung Cell Atlas (Travaglini et al., Nature 2020). Similarly, some disease-associated cell states, such as aberrant basaloid cells (ABCs) and CTHRC1^{+} fibroblasts, were annotated based on congruent expression of defining transcriptional signatures described in prior studies of pulmonary fibrosis and COPD (e.g., Habermann et al., Nat Med 2020; Adams et al., Sci Adv 2020). While inflamed non-immune cells have been less extensively described, they have been reported in certain contexts and have clear biological underpinnings. Additional references describing these inflamed cell states, including "inflamed fibroblasts," have been incorporated, and we now clarify these details in the Methods section in addition to Supplementary Table 1.

Comment 9: Results and Figures/Figure legends: Throughout the manuscript and in Figure legends, where p-values are reported, please state the name of the statistical test used to generate the p-value, as well as the sample size analyzed.

Response 9: We acknowledge the Reviewer's comment and in response, we have carefully revised the manuscript and all figure legends to include the statistical test used for each analysis, along with the corresponding sample size. Sometimes, we include this at the end of the figure legend, particularly when the test was used more than once.

Comment 10: Figure 6C: Is it possible to add confidence bands to this line plot? Also, please specify units for the y-axis.

Response 10: We thank the Reviewer for their helpful suggestion. We have revised this figure completely, added confidence bands (interquartile ranges) to the line plot in Fig. 4C,D and have clarified that the y-axis represents fold change in cell proportion compared to NS/NS Cluster 7).

Comment 11: Results/Discussion: The text notes "ME12, enriched in antiprotease SERPINA1 and inhibitors … inversely correlated with inflamed non-immune cells" It is worth further contextualizing this finding either in the Results or Discussion, as SERPINA1 is the gene coding for Alpha-1 Antitrypsin, deficiency of which is the most well-documented genetic cause of COPD and emphysema. In particular, are your findings consistent with the current understanding of AATD and its impact on risk of COPD/emphysema, or does your finding lead to new hypotheses in that regard?

Response 11: This is a really interesting point. We were also interested in this finding. What is notable is that SERPINA1 protein inversely correlates with inflamed cell states. One hypothesis is that the loss of SERPINA1 causes inflammation due to a lack of antiprotease activity and destruction. Alternatively, do the inflamed cell states promote inflammation and SERPINA1 depletion. We suspect both are possible. We have now added a sentence to the results section; "The ME12 protein module, enriched in antiprotease SERPINA1 (alpha-1 antitrypsin), was inversely associated with inflamed non-immune cell states, suggesting a reciprocal relationship in which loss of antiprotease activity may promote inflammation, while inflamed microenvironments may contribute to SERPINA1 depletion."

Comment 12: Methods – Statistics: Please address the issue of multiple comparisons, including what statistical analyses in the manuscript require these considerations, and incorporate these corrections as appropriate in the Results and corresponding interpretation.

Response 12: We have undertaken an extensive reanalysis of the entire manuscript and applied the Benjamini-Hochberg correction to all our analyses. We have also amended the Methods section to address this and other statistical issues related to the compositional analyses applied.

Minor:

Comment 13: Methods: There is a subsection labeled "Clustering of the entire LGRC dataset". Please clarify what is the relationship between LTRC and LGRC.

Response 13: The Lung Tissue Research Consortium (LTRC), established by the NHLBI, serves as a biobank providing lung tissue, blood samples, clinical data for research on lung diseases. The Lung Genomics Research Consortium (LGRC) was an extension of the LTRC that integrated genetic, molecular, and quantitative clinical phenotyping data to complement the existing LTRC clinical database. Mention of the LGRC was an error. We have removed reference to this in the manuscript.

Reviewer 3:

Authors conducted snRNAseq analysis of lung tissue from a large cohort of 141 subjects with COPD. They identified the shifts in cell composition and some pathologic cell states that correlates with clinical traits including lung function, emphysema, and symptoms. Some findings were validated through immunofluorescence and

spatial transcriptomics. The study design is comprehensive, include proteomic profiling of the ECM and blood, and spatial transcriptomics and immunofluorescence of tissues to provided cross-multi omic validations and identify plasma biomarkers reflecting pathologic lung cell states. The study incorporates multiple analyses, including spatial niche analysis, mediation analyses and cell communication networks, to reveal signaling pathways driving disease. However, the present study raised several major concerns as discussed below.

Major Comments:

Comment 1: Since the major impact of this manuscript is to map the COPD heterogeneity, to identify the shifts in cell composition and pathologic cell states that correlates with important clinical traits (i.e., lung function, emphysema, and symptoms, etc.), in this case, design to include normal lung control cohorts is necessary, the "shift" and "pathologic" states or “aberrant” cell populations can only be identified when control is present and compared. The integrated Human Lung Cell Atlas (HLCA, 37291214) or LungMAP CellRef (37516747) can be used for this purpose.

With the integration and comparison of the control lung data, authors might be able to identify pathologic cell types or states that are unique to COPD in addition to the aberrant basaloid cells (ABCs) and CTHRC1+ fibroblasts which were previously identified from pulmonary fibrosis studies. In addition, normal lungs can serve as important negative control to assess the inflammatory states in several cell types of COPD but not in the corresponding control cell types; and normal lung cell types or states do not correlate or correlate differently with clinical features.

Response 1:

We apologize for any lack of clarity. The dataset did include healthy control subjects. This is now made clear in Table 1, which we have embedded into the main text for improved clarity and referenced in response to Comment 2 above. The cohort included 13 never-smokers and 15 former or current smokers with normal spirometry. We considered integrating additional external datasets; however, we were concerned that differences in nuclear isolation techniques could introduce technical bias. That said, we have contacted the HLCA and have deposited our data there to enable future cross-cohort analyses. Importantly, all reported findings reflect changes relative to a “healthy” reference. Specifically, we compared former and current smokers to never-smokers, GOLD stages I–IV to individuals without COPD (including both never-smokers and those with normal spirometry), and individuals with radiographic emphysema to those without. These categorical comparisons were complemented by analyses of continuous clinical traits. We also performed a multi-trait clustering analysis (Figure 4A–C). Together, these analyses support that the observed changes are not limited to variation within COPD, but also reflect differences between healthy individuals and those with disease. Finally, unlike certain diseases, the distinction between healthy and diseased states in COPD can be subtle. Age, undocumented environmental exposures, imprecise smoking histories, genetic susceptibility, and other latent factors undoubtedly contribute to disease pathogenesis. As such, COPD may resemble a gradual progression associated with aging rather than a sharp dichotomy between health and disease. Collectively, we believe our dual approach—incorporating both categorical comparisons and continuous trait analyses—addresses the reviewer’s concern.

Comment 2: Aberrant cellular communities are spatially organized within distinct neighborhoods, suggesting a critical role for cellular microenvironments in disease progression. This is an important finding of the paper. However, the figures from spatial transcriptome analysis are in very low resolution and cannot serve as strong supporting evidence. Xenium platform is capable to provide high resolution to detect marker expression at subcellular levels, together with quantitively measure of the cell selective markers expression and cell composition in representative niches (zoom in view) would be a better way to visualize the data.

Response 2: We agree with the comments of the reviewers and have revised the manuscript to improve the visualization of the data. Consequently, we have redone Figure 5. Additionally, Extended Data Figures 5,6,and 7 all provide different visualization and representations of the data.

Comment 3: Correlation of cell proportions with clinical features derived major findings in this work. However, a simple correlation analysis was used without considering the impact of covariates, such as age, sex, height, BMI, etc. on the analysis.

Response 3: We have repeated all analyses and now adjust for age, sex, and BMI in all models. We did not adjust for height, as it is already incorporated into the calculation of predicted spirometric values. We chose not

to adjust for race, as doing so may obscure the influence of racial disparities on disease biology (PMID: 37789703). In our view, the role of race in shaping disease manifestations warrants a dedicated investigation and is outside the scope of the present study. Similarly, additional variables such as lobe of origin, medication use, and comorbidities represent important sources of variation, but each introduces complex biological and clinical questions that are best addressed in focused, independent studies. We agree that these are important considerations, and our current approach lays the groundwork for such future analyses**.**

Comment 4: Authors revealed cellular pathways and signaling networks associated with clinical outcomes and inflamed cell states via connectome analysis. Again, it is important to show the comparison with normal controls. Does any of the ligand-receptor or signaling pathway prediction can be cross validated via proteomic profiling and spatial transcriptomics data?

Response 4: We have taken an extensive analysis of the Xenium data to validate findings through spatial transcriptomics This is now included in Figure 7E,F. with methods fully described in the text (Methods section). Briefly, we inferred ligand–receptor signaling using NicheNet, applying two complementary strategies. In the receiver-oriented approach, structural cells, alveolar macrophages (AM), and interstitial macrophages (IM) were analyzed individually as receivers, and differentially expressed genes (DEGs) between inflamed vs. non-inflamed (or between specified subtypes) were identified with Seurat's *FindMarkers* and used as NicheNet targets, retaining the top 30 ligands by AUPR. In the sender-oriented approach, the same structural populations were designated as senders, with upregulated ligands identified between inflamed vs. non-inflamed senders, and target genes defined from GLMMs (iDESC) linking sender frequency to receiver expression. Predicted interactions were spatially validated using matched 10x Xenium data by computing bivariate local Moran's I statistics on curated CellChat ligand–receptor pairs (excluding ECM), assessing either ligand–receptor co-localization or ligand–receiver cell-type proximity when receptor expression was absent. Hotspots ($z > 2$) were compared between inflamed and non-inflamed groups by prevalence (two-proportion z-test) and intensity (Wilcoxon), with Benjamini–Hochberg correction; interactions significant in both tests (FDR < 0.05) were deemed spatially enriched.

Comment 5: Data available: data has not submitted to GEO or dbGAP.

Response 5: We acknowledge the Reviewer's comment and believe this is important. Raw and processed data have been deposited in the NIH's Sequence Read Archive (SRA) database. They gave us the BioProject ID: PRJNA1301244. Additionally, we had extensive conversations with representatives from NHLBI. Because the clinical data was generated by a NIH funded study, we are not allowed to report the clinical data for each individual ourselves. However, investigators can access these data by applying for permission from BioLINCC, the NHLBI biorepository.

Minor Points:

Comment 6: Most of the figures were mixed with clear panels and fuzzy panels, some are unreadable (e.g., 1D, 3J, 4A, 4I, 7D-I, 9D, 9F and many supplementary panels)

Response 6: We appreciate the Reviewer's comment. All figures have now been remade using Adobe Illustrator and therefore should no longer suffer from the loss of resolution that can occur with .ppt > .pdf conversion.

Comment 7: Fig.1C: notes left in umap "change cluster number to final cell annotation"

Response 7: We thank the Reviewer for pointing this out. The annotation notes have been removed from **Fig. 1C**, and the figure has been updated to reflect the finalized cell cluster numbers.

Comment 8: The correlation between cell states and clinic severity measures are moderate, nothing above 0.5, which is commonly considered as the threshold for significant correlation.

Response 8:

We appreciate the reviewer's observation. We acknowledge that the clinical severity measurements are moderate. However, given COPD's biological complexity, moderate effect sizes should be expected. For example, recent reviews of the COPD biomarker literature highlight that most published blood-based markers (e.g., CRP, IL-6, fibrinogen, SP-D, CC16, sRAGE) show only modest correlations with lung function or

emphysema (typically r ≤ 0.3–0.4) and explain <10 % of outcome variance, largely because of assay variability, heterogeneous disease mechanisms, and wide overlap between patients and controls (Stockley et al., Am J Respir Crit Care Med 2019) Additionally, what is often less discussed is the lack of precision of common tools, like spirometry, to accurately assess the biology of disease, and anchoring on clinical measurements of disease only obscures this heterogeneity. In fact, the authors of this review comment that "*the complexity of biomarker specificity without sufficient clinical phenotype and endotype information contributes to problems of interpretation. A strategic change is needed to develop useful COPD biomarkers; this includes focusing on endotype biomarkers within specific clinical phenotypes."* Our data here defines those endotypic features, and because they are, by definition, not consistent across all subjects we interpret these effect sizes as biologically plausible and clinically relevant signals.

Comment 9: Fig.3B, 3G: subtypes distinction is subtle. No selective markers to distinct them; not sure if these transitional states were only in COPD but not in control.

Response 9: Subtype distinction here is indeed subtle, and we agree that this represents a cell state, and in fact, likely represents a continuum of disease. However, all these cell states have been previously described. We now write in the methods section. *Emergent cell states were defined by distinct transcriptional signatures and supported by gene set enrichment analysis; inflammatory non-immune cells were defined by their enrichment for inflammatory genes/pathways or described in results, while AT2$_S$,[14,52] AT2$_{SCGB3A2}$,[11-13] ABC,[9,10] fibroblast populations,[12,15,24,32,53] , gCap$_S$[22,54,55] and IM$_{CHIT1}$[56] were annotated based on unique transcriptomic signatures and supported by prior literature.*

Comment 10: Fig 4I and J, the y axis values are confusing. They were labeled as "cell proportions (log2 scale)". Proportion values should be within 0 and 1, and thus log2 scale will become negative values. However, in the y axis of Fig 4I and J, all log2 cell proportion values were positive.

Response 10: We appreciate the reviewer's astute eye, and the y-axis labels are now corrected. We did not perform a log transformation of the proportion; we performed a log transformation of the percentage (0-100). This is now made clear in the figure legends.

Comment 11: "COL15A1+ systemic peribronchial endothelial cells increased with emphysema severity (P=0.003) (Fig. 2E)." systemic peribronchial endothelial cells were increased with emphysema stage in Fig 2b & 2C, not Fig.2E

Response 11: We have now corrected this mistake.

Comment 12: EXT Fig 9 shows the umaps of Xenium by lineages, can you make a complete umap with marker genes support?

Response 12: We have generated a comprehensive UMAP that overlays marker gene expression across all major lineages in the Xenium dataset. We now include this in the manuscript as **Extended Data Panel 5**.